Portage de ePANAM sur la grille de calcul

T. Tung DOAN (1), Najwa TAIB (2), H.Quang NGUYEN (1), Jean-Christophe CHARVY, Gisele BRONNER (2), Vincent BRETON (3), Didier DEBROAS (2), Engelbert MEPHU (4)

(1) {dttung, nhquang}@ifi.hut.edu.vn, Institut de la Francophonie pour l'Informatique – 42 Ta Quang Buu, Hanoi, Vietnam
(2) {najwa.taib, j-christophe.charvy, gibronne, didier.debroas}@univ-bpclermont.fr, LGME, Université Blaise Pascal, 24 avenue des Landais, BP 80026, 63171 Aubière Cedex
(3) breton@clermont.in2p3.fr, Laboratoire de Physique Corpusculaire, CNRS/IN2P3, 24 avenue des Landais, BP 10448, F-63000 Clermont-Ferrand, France
(4) Engelbert.MEPHU_NGUIFO@univ-bpclermont.fr, ISIMA, Campus des Cézeaux, BP 10125, 63173 Aubière Cedex

Overview

We present in this paper the gridification of ePANAM - a pipeline dedicated to analyze massive sequencing results that automatically affiliates sequences from SSU rRNA amplicons and build phylogenetic trees of very large numbers of sequences. The porting of ePANAM is based on WISDOM Production Environment – a grid middleware designed as an experiment management environment for handling jobs and data on the grid. The grid version of ePANAM improves significantly the run-time, especially on huge data of Next Generation Sequences (NGS).

Enjeux scientifiques, besoin de la grille

PANAM [1] est un pipeline dédié à l'analyse de séquences environnementales issues du séquençage massif de nouvelle génération (NGS). Il permet l'annotation phylogénétique automatisée de séquences d'amplicons de la petite sous-unité de l'ARNr. Ce pipeline doit être proposé à la communauté scientifique et intégré à ePANAM, une ressource Web pour l'analyse de la diversité microbienne. ePANAM permet de faire un contrôle sur la qualité des séquences, de calculer des indices de richesse et de diversité de comparer les écosystèmes, d'attribuer une taxonomie aux séquences expérimentales et de décrire les clades environnementaux d'intérêt. Ce pipeline est basé sur des outils disponibles publiquement (USEARCH [2], HMMER [3] et FASTTREE [4]).
Sur un processeur Intel à 2 GHz (R) Xeon (R) CPU 32 bits et 24 Go de RAM, ePANAM permet de réaliser la phylogénie de 1000 séquences de 200 pb (paires de base) en environ 20 minutes. Pour des séquences de 400 pb générées par les pyroséquenceurs actuels, le temps d'exécution varie de 24 minutes pour 5.000 séquences à 6 jours et 14 heures pour 1.000.000 de séquences. Enfin, l'analyse phylogénétique d'un million de séquences complètes est réalisée en 16 jours. Ainsi, ePANAM représente une première réponse au problème lié à l'analyse phylogénétique d'amplicons issus des NGS par rapport aux méthodes existantes. Cependant, bien que le temps d'exécution soit adapté à une utilisation individuelle, il reste prohibitif pour que cette approche devienne une ressource web de premier plan pour la communauté mondiale des microbiologistes. Ainsi, les performances de ePANAM seraient améliorées en utilisant des approches de parallélisation et des technologies comme la grille de calcul.

Développements, déploiement sur la grille

Le traitement des séquences par ePANAM est réalisé en deux étapes : la première étape compare les séquences à une base de référence constituée de séquences extraite de SILVA [5] avec USEARCH, ensuite les séquences expérimentales sont associées à des groupes phylétiques prédéfinis, constitués à partir des séquences de référence. Dans la deuxième étape, les séquences expérimentales sont alignées contre les profils d'alignement des

séquences de références auxquelles elles sont associées. Un arbre phylogénétique contenant les séquences de référence et les séquences expérimentales est généré pour chaque groupe phylétique, et une affiliation taxonomique est inférée pour chaque séquence expérimentale selon sa position dans l'arbre.
Le déploiement de ePANAM sur la grille est basé sur « WISDOM Production Environment » [6]. Le pipeline de ePANAM est divisé en 4 services de WPE. Le service *g-panam-trim* adapte les profils d'alignement des séquences de références à la région amplifiée des séquences expérimentales avant de lancer le traitement phylogénétique. Le service *g-panam-split* qui trie les séquences selon le groupe phylétique inféré par USEARCH et les affecte aux différents groupes phylétiques. Ensuite, le service *g-panam-alnphy* aligne des séquences par un outil d'alignement (HMMER), avant de construire les phylogénies par FASTTREE. Cette phase est parallélisée selon le nombre des fichiers récupérés à la fin du service précédant (*g-panam-split*). Finalement, le service *g-panam-clade* parcourt les arbres, assigne une taxonomie pour chaque séquence et infère les clades putatifs.

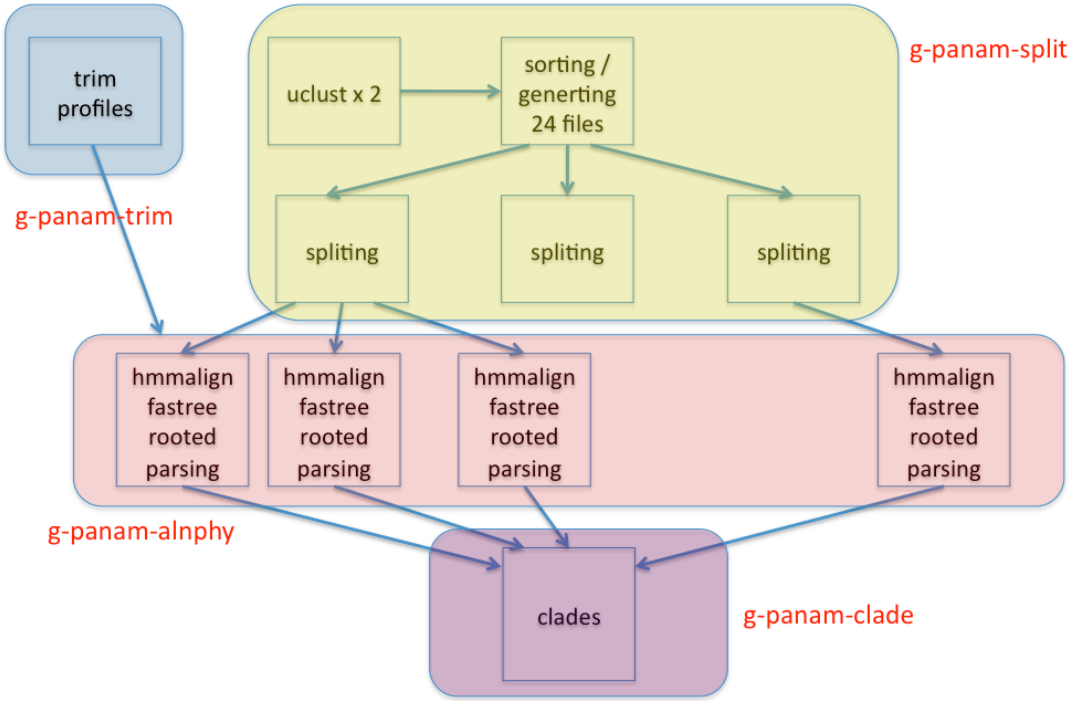

Figure 1 – Porter ePANAM sur la grille

Résultats scientifiques
La figure 2 présente les temps d'exécution de ePANAM avec un jeu de 250.000 séquences de 400 pb pour deux méthodes de parcours de phylogénies : LCA (*lowest common ancestor*) et NN (*nearest neighbor*) sur la grille. Le temps de traitement total de ePANAM sur la grille varie de 8h13' à 9h43' (méthode NN). Sur un ordinateur personnel, ePANAM prend 11h49' avec la même méthode. Le temps de la tâche *g-panam-split* varie de 1h46' à 3h15'. Pour la tâche de l'alignement et de la phylogénie, la méthode LCA ne prend que 20' maximum cependant, la méthode NN dure de 1h8' à 1h19'. Dans ce test, la tâche *g-panam-split* génère 25 instances de la tâche *g-panam-alnphy*.

Workflow	Total	Split	Mean AlnPhy	Clade
Test70s25lca	3h10	2h39	0h10	0h14
Test107s25LCA	5h32	3h15	0h20	2h17
Test109s25LCA	2h36	1h46	0h4	0h50
Test80s25NN	9h7	1h44	1h8	3h26
Test81s25NN	8h13	2h13	1h13	5h59
Test79s25NN	9h43	1h53	1h19	7h50

ePANAM (250K, NN): 11h49

Figure 2 – Test de ePANAM sur la grille (250000 séquences de 400 pb)

La figure 3 présente les durées d'exécution sur 500.000 séquences de 400 pb.
En comparaison avec le traitement sur un ordinateur personnel (38h25'), le temps de traitement est beaucoup plus réduit sur la grille, du même ordre de grandeur que ceux obtenus pour le traitement de 250.000 sequences. Dans ce test, la tâche *g-panam-split* génère 31 instances de la tâche *g-panam-alnphy*.

Workflow	Total	Split	Mean AlnPhy	Clade
Test88s50LCA	6h22	2h33	0h22	3h48
Test86s50NN	7h52	2h11	4h7	5h39
Test98s50NN	9h49	2h15	3h58	6h29
Test100s50NN	10h17	1h42	2h33	7h34

ePANAM (500K, NN): 38h25

Figure 2 – Test de ePANAM sur la grille (500000 séquences de lecture 400 pb)

Perspectives

Le premier portage de ePANAM sur la grille a montré l'intérêt de cette approche sur de grands jeux de données d'amplicons, confirmant l'efficacité de la grille pour la résolution de problèmes biologiques similaires. Il reste à résoudre le problème de l'expiration de proxy sur la grille dans les prochains tests avec les tailles de données plus grandes. Pour mettre en application cette approche comme une ressource de la communauté mondiale des microbiologistes, il faut implémenter un moteur de workflow pour contrôler le pipeline.

Références

[1]. Taib N, Bronner G, Debroas D (2010). Annotation phylogénétique de sequences du gène codant pour l'ARNr 18S généré par pyrosequençage (PANAM). Rencontres ALPHY – 2-3 Février - Marseille.
[2]. Edgar RC (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26: 2460-2461.
[3]. Eddy SR (1998). Profile hidden Markov models. Bioinformatics 14: 755- 63.
[4]. Price MN, Dehal PS, Arkin PA (2009). FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. Mol. Biol. Evol. 26(7): 1641-1650.
[5]. Pruesse E, Quast C, Knittel K, Fuchs B, Ludwig W, et al (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucl. Acids Res. 35: 7188-7196.
[6]. V. Breton, A. L. D. Costa, P. D. Vlieger, L. Maigne, D. Sarramia, Y. Kim, D. Kim, H. Q. Nguyen, T. Solomonides, and Y. Wu (2009), *Innovative in silico approaches to address avian flu using grid technology*, Infectious Disorders Drug Targets, 9(3):358-65.